ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Title goes here

Master Thesis

Robin Bosshard, 16-915-399

October 16, 2023

Supervisors: Prof. Dr. Fernando Perez-Cruz, Dr. Eliza Harris, Lili Gasser (SDSC)
Dr. Kasia Arturi (Eawag)

Department of Computer Science, ETH Zürich

Contents

Chapter 1

Introduction

1.1 The Challenge of Environmental Pollution

Over the past few decades, the upsurge in environmental pollution by chemical compounds has been driven by industrial processes, agricultural methods, our consumerism and various other contributing factors. Although these chemicals are integral for many products and have the potential to improve our comfort of modern society, they can also pose risks and adversely affect both our health and the environment, either acutely or chronically. Toxic substances threaten wildlife but also make our air, soil and finally our drinking water and food supply less safe.

The EU maintains comprehensive chemical regulations, however, it is anticipated that global chemicals production will double by 2030 [1]. Moreover, the widespread utilization of chemicals, including their inclusion in consumer goods, is expected to expand further. Even though there are over 275 million known chemical compounds registered by the *Chemical Abstracts Service* [2], merely a tiny fraction of them undergo close monitoring via target analytical approaches and even less is known about their toxicity profiles and negative health effects on our organsims. Table 1.1 provides an overview of omnipresent water pollutants.

Building upon the *European Green Deal* [4] and the *8th Environment Action Programme*, which guides European environmental policy until 2030, reinforces the EU's goal of sustainable living within planetary limits, with a vision extending to 2050. One of its key objectives is a zero-pollution commitment, covering air, water, and soil, prioritizing the well-being of EU citizens. In particular, the *European Commission* published a sustainability-focused chemicals strategy, aligning with the EU's zero-pollution ambition with one of the objectives to minimize concerning substances by either substituting or phasing them out wherever feasible [5]. Consequently, the urgent need to monitor and effectively assess the hazards associated with the daily entering

of thousands of poorly understood chemicals into our environment becomes increasingly evident.

1.2 The Imperative for Prioritization and Toxicity Assessment

Contemporary analytical techniques, particularly *high-resolution mass spectrometry (HRMS/MS)*, are gaining significance across various domains such as metabolomics, drug discovery, forensics, environmental science, and toxicology. *Nontarget HRMS/MS* has improved the ability to detect emerging compounds in environmental samples, often with unknown toxicity profiles. These compounds are assessed based on factors such as abundance and fragmentation data. However, the task of identifying compounds and understanding their toxicity continues to be demanding in terms of resources and time. Additionally, the scarcity of thoroughly characterized reference substances for comparison when studying unknown compounds adds complexity, making it difficult to achieve comprehensive elucidation. Traditionally, the prioritization of unidentified compounds rely on signal intensity as a guiding metric. Unfortunately, this approach falls short in delivering an accurate assessment of environmental exposures, as it tends to overlook the crucial toxicological dimension. As a result, substances with the potential for severe ecological consequences, such as endocrine-disrupting compounds, often go undetected because of their low abundance, even though they exhibit high levels of toxicity. Hence, a pressing requirement exists for alternative approaches to prioritize unidentified *NTS HRMS/MS* signals based on their hazard potential, which can better incorporate considerations of toxicity and ecological consequences. Figure 1.1 illustrates the non-target screening with *HRMS/MS* technique and the novel prioritization approach.

1.3 Unlocking the Potential of High-Throughput Screening and Machine Learning in Toxicity Prediction

In the past few years, the use of machine learning methods has emerged as a transformative force in the field of *in vitro* toxicology, particularly in the realm of high-throughput toxicity prediction. High-throughput screening (HTS) has revolutionized the way we assess toxicity by allowing thousands of *in vitro* bioassays to be conducted efficiently. This high-throughput approach, coupled with advancements in robotics and automated analysis, has generated large volumes of toxicity data, paving the way for more comprehensive assessments of chemical compounds. Alongside the rise of machine learning, this advancement has facilitated the creation of predictive models capable of forecasting compound toxicity based on their chemical structure [7]. These

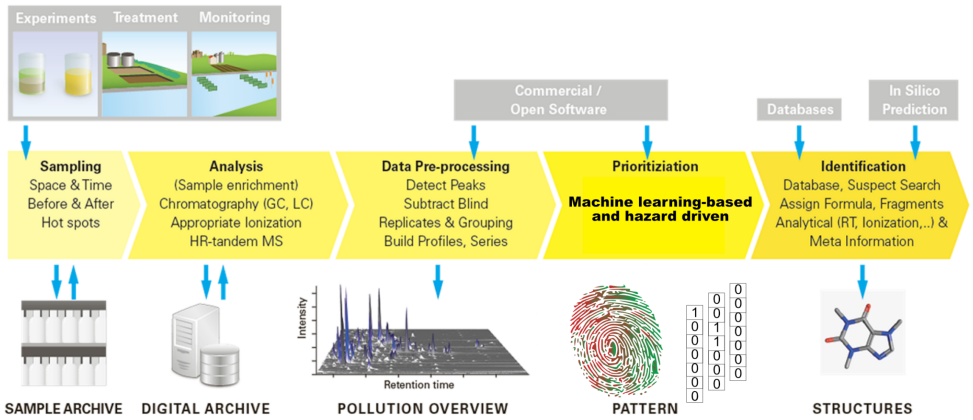
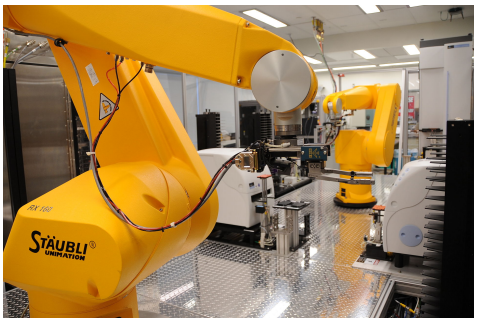
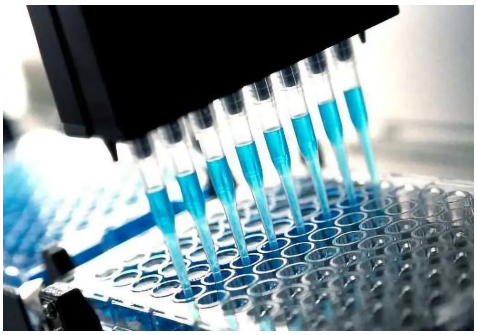


Figure 1.1: Figure 1 adapted (with modified Prioritization step) from Hollender et al. [6]: Nontarget screening with high resolution mass spectrometry in the environment: Ready to go?

models can be trained on extensive datasets containing well-documented toxicity information, allowing them to learn the underlying patterns and relationships between chemical structures and target toxicity. Once trained, these models can predict the toxicity of new compounds, even if they have not undergone laboratory testing. This approach holds the potential to significantly reduce the time and cost associated with early-stage toxicity pre-assessment and plays a crucial role in prioritizing compounds for further in-depth testing.



(a) A robot arm retrieves assay plates from incubators and places them at compound transfer stations or hands them off to another arm that services liquid dispensers or plate readers. Efforts in the automation, miniaturization and the readout technologies have enabled the growth of HTS. Image obtained from [8]



(b) Modern microtitre assay plates consist of multiples of 96 wells, which are either prepared in the laboratory or acquired commercially from stock plates. These wells are filled with a dilution solvent, such as DMSO, along with any other chemical compounds intended for analysis. Image obtained from [9]

Figure 1.2: High-Throughput Screening (HTS)

1.4 MLinvitroTox: A Novel Approach

In response to the pressing need for a more hazard-driven and comprehensive assessment of environmental contaminants, *Arturi et al.* introduced *MLinvitroTox* [10], an innovative machine learning framework. The primary objective of this thesis is to collaborate with the authors to further enhance and advance this framework. *MLinvitroTox* leverages molecular fingerprints extracted from fragmentation spectra[1], marking a fundamental shift in how we forecast the toxicity of the myriad unidentified *HRMS/MS* features. While traditional *QSAR* models predict bioactivities based on molecular fingerprints derived from chemical structures, *MLinvitroTox* was trained with supervised classification models on molcular fingerprints from chemical structures but is applied to molecular fingerprints generated from experimentally measured *MS2* spectra using *SIRIUS* and *CSI:FingerID* [11]. *SIRIUS* is a software package for annotating small molecules from nontarget HRMS/MS data, while *CSI:FingerID* is a machine-learning tool employed by *SIRIUS* to predict molecular fingerprints from fragmentation spectra. *MLinvitroTox* leverages streamlined machine learning techniques to predict the compounds bioactivity, respectively toxicity, ensuring a broad toxicological coverage encompassing over 400 target-specific and 70 cytotoxic endpoints, sourced from *ToxCast/Tox21* data. Subsequently, the toxicity predictions generated by the framework are employed to prioritize compounds, with the flexibility to emphasize specific aspects of toxicity profiles tailored to individual preferences. This prioritization strategy facilitates more streamlined and thorough evaluations of environmental contaminants, enhancing a more hazard-driven risk assessment.

1.5 Objectives and Significance

The primary goal of this thesis is to contribute to the development of an efficient framework for predicting compound toxicity across multiple endpoints. The ultimate output of the pipeline are the toxicity fingerprints that encode the predicted toxcity from *HRMS* environmental samples for the relevant endpoints of interest. The generated toxicity fingerprints will provide valuable insights for the prioritization process in identifying most hazardous compounds found in environmental samples, ultimately contributing to the preservation of ecosystems and our health. The framework aims to develop a custom processesing and curation of structural and toxicological data to address challenges from modeling heterogenuous and imbalanced data sets.

[1] also termed as *Tandem mass spectrometry* or *MS/MS* or *MS2*

1.6 Thesis Structure

The initial chapters lay the groundwork by providing essential background information and summarizing related work. As we progress through the subsequent chapters, we will explore the materials and methods employed, providing insights into the technical intricacies of preparing ToxCast/Tox21 toxicity data and transforming them into appropriate inputs for our machine learning pipeline. This foundational work will serve as the cornerstone for the forthcoming chapters, where we will showcase the potential of *MLinvitroTox*. Additionally, will also demonstrate the framework's effectiveness through validation using real-world mass spectral data from *MassBank* [12] and discuss about the implications of our research.

Origin/Usage	Class	Examples	Related Issues
Industrial Chemicals	Solvents	Tetrachloro-methane	Drinking-water-quality
	Intermediates	Methyl-t-butylether	Drinking-water-quality
	Petrochemicals	BTEX (benzene, toluene, xylene)	Cancer
Industrial Products	Additives	Phthalates	Endocrine disruptors
	Lubricants	PCBs	Biomagnification
	Flame Retar-dants	PBDEs	
Consumer Products	Detergents	Nonylphenol ethoxylates	Endocrine effects
	Pharmaceuticals	Antibiotics	Bacterial resis-tance
	Hormones	Ethinyl estradiol	Feminization of fish
Biocides	Pesticides	DDT	Toxic effects and persistent metabolites
	Nonagricultural biocides	Tributyltin	Endocrine effects
Geogenic & Natural Chemicals	Heavy Metals	Lead, cadmium, mercury	Organ damage
	Inorganics	Arsenic, sele-nium, fluoride	Drinking-water-quality
	Taste and Odor	Geosmin	
	Human Hor-mones	Estradiol	Feminization of fish
Disinfection & Oxidation	Disinfection by-products	Haloacetic acids, Bromate	Drinking-water-quality
Transformation Products	Metabolites from all above	Metabolites of perfluorinated compounds	Bioaccumulation
		Chloroacetanilide herbicide metabolites	Drinking-water-quality

Table 1.1: Table 2 adapted from [3]. Examples of ubiquitous water pollutants.

Chapter 2

Background

This chapter provides information essential for comprehending the subsequent sections of this thesis by introducing the challenges and the evolving trends in toxicity testing.

2.1 Toxicity Testing: From In Vitro Assays and Molecular Fingerprints to Predictive Models and Beyond

With the ever-growing amount of chemical compounds entering our environment, conducting experiments on all these compounds using traditional methods have constraints related to expense and time, and ethical considerations regarding animal trials.

In 2007, the *U.S. National Academy of Sciences* introduced a visionary perspective and published a landmark report, titled as *Toxicity Testing in the 21st Century: Vision and Strategy*. This report promoted a transition from conventional, resource-consuming animal-based *in vivo* tests to efficient high-throughput *in vitro* pathway assays on cells.

This transition paved the way for the realm of high-throughput screening (HTS), where a multitude of *in vitro*bioassays can be executed, complementing and improving chemical screening. This transformation is made possible by advancements in robotics, data processing, and automated analysis. As a result, this synergy has led to the generation of extensive toxicity datasets like ToxCast and Tox21.

HTS datasets, including *ToxCast* and other sources, have opened the door to promising applications of machine learning in predictive computational toxicology. PThese predictive models can be developed to screen environmental chemicals with limited toxicity data availability, allowing for the prioritization of further testing efforts. Such models often forecast toxicity using *Quantitative Structure-Activity Relationships (QSARs)*, which are based

on chemical structures encoded as descriptors, such as molecular fingerprints. 1D-Molecular fingerprints encode compound molecules as fixed-length binary vectors, denoting the presence (1) or absence (0) of specific substructures or functional groups.

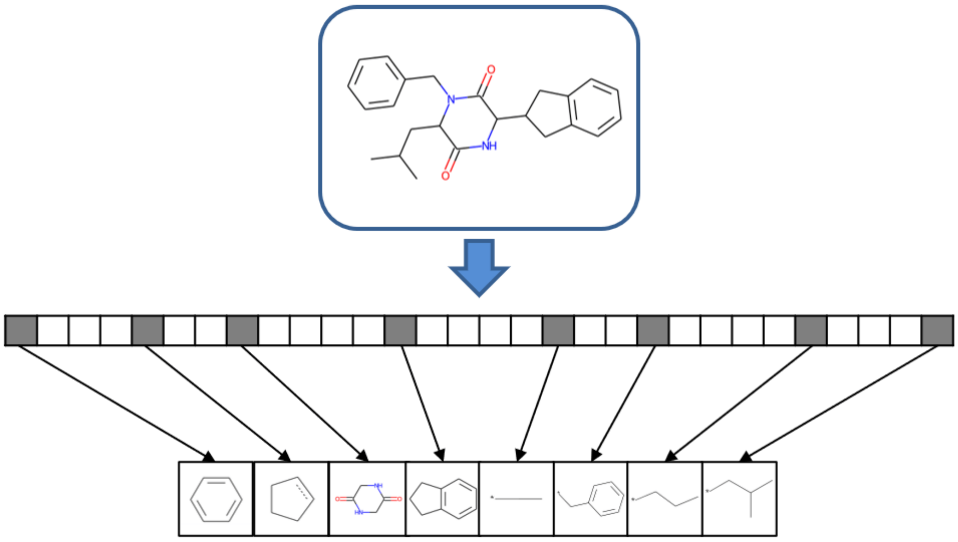


Figure 2.1: Figure 1 adapted from Janela et al (2022) [13]. Schematic molecular fingerprint. Each bit position accounts for the presence or absence of a specific structural fragment. Bit positions are set on (set to 1, gray) if the substructure is present in a molecule, or set off (set to 0, white) if it is absent.

The utilization of molecular fingerprints for in vitro toxicity prediction is based on the assumption that molecular toxic effects result from relatively straightforward interactions between distinct chemical components and receptors during a *molecular initiating event (MEI)*. On a larger biological scale, the *MEI* can set a sequential chain of causally linked *key events (KE)* in motion. This occurs at different levels of biological organisation from within cells to potentially culminating in an *adverse outcome pathway (AOP)* at the organ or organism level, as depicted in Figure 2.2. The mechanistic information captured in *AOPs* reveal how chemicals or other stressors cause harm, offering insights into disrupted biological processes, potential intervention points but also guide regulatory decisions on next generation risk assessment and toxicity testing. The *AOP* framework is an analytical construct that allows an activity mapping from the presence or absence of certain molecular substructures encoded in chemical descriptors to the target mechanistic toxicity. Finally, when monitoring disruptions in toxicity pathways, *physiologically based pharmacokinetic (PBPK)* models can be leveraged to extrapolate *in vitro* findings to human blood and tissue concentrations [14].

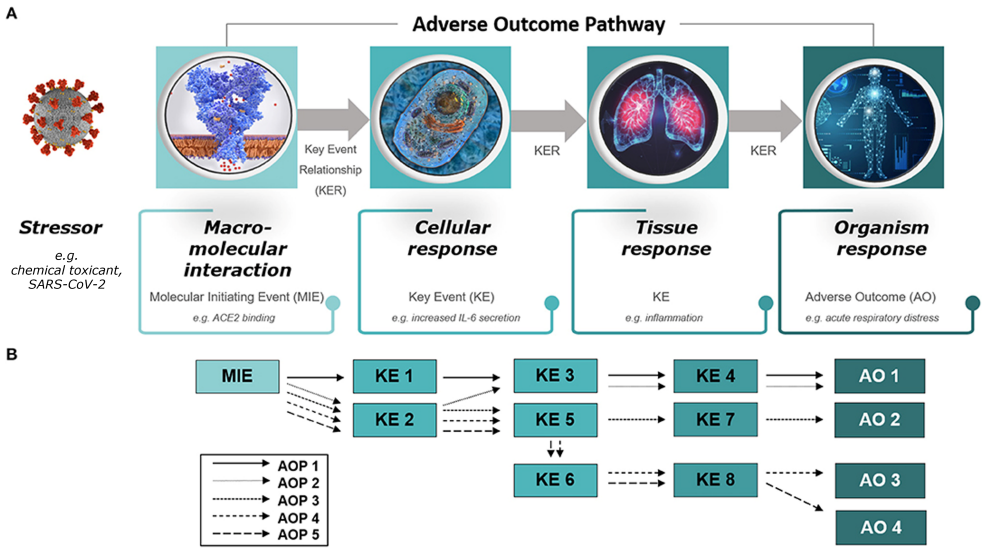


Figure 2.2: Figure 1 adapted from [15]: Diagram of (A) an adverse outcome pathway (AOP) and (B) an AOP network. (A) An AOP starts with a molecular initiating event (MIE), followed by a series of key events (KEs) on different levels of biological organization (cellular, tissue, organ) and ends with an adverse outcome (AO) in an organism. The stressor is not part of the AOP itself.

2.2 Chemical Target Toxicity vs. Cytotoxicity

Intuitively, we expect increasing chemical concentrations to result in increasing bioactivity. However, this is not always the case. At higher doses the chemical can become cytotoxic leading to dying cells. In consequently, a reduction in bioactivity can occur, e.g., the activation of the reporter gene decreases.

Chemical toxicity can manifest in diverse ways, falling into two major categories [16]:

- **Specific toxicity** is the result of a chemical's interaction and disruption of a specific biomolecular target or pathway, such as a receptor agonist/antagonist effect or enzyme activation/inhibition.
- **Cytotoxicity and cell stress** is the generalized disruption of the cellular machinery. Cell-disruptive processes encompass various mechanisms, such as protein, DNA, or lipid reactivity, or processes like apoptosis, oxidative stress responses or mitochondrial disruption. Cell viability can be evaluated either individually or concurrently. One approach is to assess it by calculating the proportion of live cells in a population, employing a fluorescent dye that specifically enters living cells. This dye remains incapable of permeating the membranes of deceased cells, resulting in fluorescence intensity directly correlating with cell viability.

It is common for compounds to exhibit target bioactivity within a limited concentration range, which coincides with a non-specific activation response in the presence of cell stress and cytotoxicity. Figure 2.3 illustrates the interference between specific toxicity and cytotoxicity.

A related phenomenon is referred to as the *cytotoxicity burst* [16],is observed, where, for example, reporter genes may be induced non-specifically near the point of cell death [17]. The *ToxCast* pipeline is designed to minimize false negatives by adopting an inclusive risk assessment approach. However, due to interference processes, the reliability of reported activities becomes uncertain, and false positives can occur without a thorough cytotoxicity evaluation. While some of the assay activity within this concentration range may indeed result from chemical interactions with the intended assay target, another portion does not and needs to be taken into account.

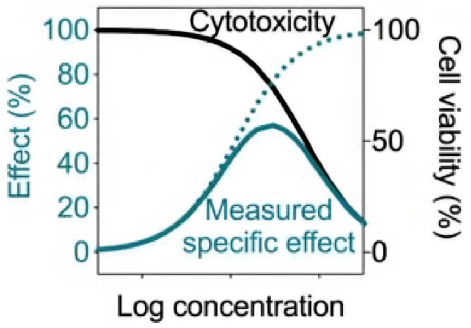


Figure 2.3: Figure 7.8 from Escher at al. [17]: Bioanalytical Tools in Water Quality Assessment: Second Edition. Example of a bioassay response with cytotoxicity interference. The dotted line shows the theoretical effect but due to cytotoxicity (black line is cell viability), the measured effect has an inverted U-shape. The measured effect can addionally be confounded and intensified by the cytotoxicity burst, where even an exponential shape is likely for the gaining part. In this case, the effect should be only evaluated up to some concentration.

Chapter 3

Related work

A recent review highlights the proliferation of research employing invitroDB since 2006, encompassing topics such as assessing chemical toxicity, identifying contaminants for environmental monitoring, and computational toxicity forecasting. The majority of ML applications based on invitroDB have predominantly concentrated on specific target endpoints and cytotoxicity. Notably, research has extensively covered adverse outcomes related to endocrine receptor systems, including androgen and estrogen receptors, alongside areas such as carcinogenicity, hepatic steatosis, hepatotoxicity, immunotoxicity, developmental toxicity, neurotoxicity, and cardiotoxicity.

Typically, various mathematical models or curve shapes are employed to analyze the data for the best fit. Several commercial tools and open-source libraries are available for this purpose. One widely used system for managing high-throughput screening (HTS) concentration-response data is tcpl (ToxCast Pipeline), but also.

Compared to similar efforts in the field where ecotoxicity was predicted from MS2 based on in vivo data, in the current work, the invitroDB toxicity database was used to train supervised classification models for hundreds of available toxicity endpoints

In their systematic investigation using Tox21 data, Wu et al. (2021) explored the impact of various modeling approaches and chemical features on predictive toxicology, with a focus on model performance and explainability trade-offs. The study found that the assay endpoint from the Tox21 data being predicted was the most significant factor influencing model performance. Endpoints with higher predictability, characterized by lower data imbalance and larger datasets, performed well regardless of the modeling approach or molecular representation. For less predictable endpoints, simpler models like Linear Regression performed similarly to complex ones, prioritizing both predictivity and interpretability. Moreover this study suggests consensus

modeling and multi-task learning to enhance predictability and model performance across endpoints. In this thesis, we set the goal to not overlook simpler models due to their higher interpretability and comparable performance. As suggested we do not further investigate on the differnt molecular representations and use a fixed compilation of molecolar fingerprints[1] as initial input features. We incorporated in our studies a form of consensus modeling to consolidate predictability and multi-task learning to improve model performance across different endpoints.

[1] SIRIUS

Chapter 4

Material and Methods

4.1 InvitroDB v4.1

The most recent release of the *ToxCast's (Toxicity Forecaster)* database[1], referred to as *invitroDBv4.1*, serves as a source of an extensive collection of *HTS* targeted bioactivity data. This database encompasses information on a total of 10 196 compounds, selectively screened across 1485 assay endpoints. The assays utilize a range of technologies to assess the impact of chemical exposure on a wide array of biological targets, including individual proteins and cellular processes such as mitochondrial health, developmental processes and nuclear receptor signaling.

This resource originated from the collaboration of two prominent institutions: the *United States Environmental Protection Agency*[2] *(EPA)* through its *ToxCast* program and the *National Institutes of Health*[3] *(NIH)* via the *Tox21*[4] initiative. Utilizing data gathered from various research laboratories, this relational database is publicly available and can be downloaded[5] by visiting the official *ToxCast* website.

4.2 tcpl v3.0

The *tcpl*[6] package, written in `R`, offers a comprehensive suite of tools for managing HTS data, provides reproducible concentration-response modeling and populates the `MYSQL` database, *invitrodb*. The multiple-concentration screening paradigm intends to pinpoint the activity of compounds, while also

[1] released on September 21, 2023
[2] https://www.epa.gov
[3] https://ntp.niehs.nih.gov/whatwestudy/tox21
[4] https://tox21.gov/
[5] https://www.epa.gov/chemical-research/exploring-toxcast-data
[6] https://github.com/USEPA/CompTox-ToxCast-tcpl

estimating their efficacy and potency. The concentration-response modeling procedure also addresses outlier robustness and signal loss due to cytotoxicity. In Chapter 4, the `Python` re-implementation *pytcpl* of the fundamental components of the ToxCast pipeline *tcpl* is introduced but at this point the essential elements are layed out that underpin the entire pipeline.

Each compound-assay dataset involves the collection of the respective *concentration-response series (CRS)* denoted as S_{ij}, representing compound c_j in assay endpoint a_i. A CRS is represented as a set of concentration-response pairs:

$$S = \{(conc_1, resp_1), (conc_2, resp_2), \ldots, (conc_{n_{\text{datapoints}_{i,j}}}, resp_{n_{\text{datapoints}_{i,j}}})\}$$

where $n_{\text{datapoints}_{i,j}}$ varies based on the number of concentrations tested for compound c_j in the assay endpoint a_i.

The concentration-response pairs can be retrieved by combining tables *mc0*, *mc1*, and *mc3* from *invitroDBv4.1*, representing the raw data. Esential sample information, including well type and indices from the assay well-plate is also collected. The concentrations are transformed to the logarithmic scale having the unit μM (micromolar), while the responses are control well-normalized to either fold-induction or percent-of-control activity. A control well is a set of sample wells that serve as a baseline for comparison to the impact of the treated samples. Control wells typically contain untreated samples or samples with a known, non-toxic response. The control wells are used to normalize the treated samples to account for any background noise or variability in the assay. There are two methods for analyzing raw assay results that will impact the analysis of the background distribution [18]:

a. **Fold Induction:** Fold induction is a measure used to quantify how much a certain parameter (e.g., gene expression or protein activity) has changed in response to a treatment compared to its baseline level. For example, if a gene is expressed five times higher in a treated sample compared to the control, the fold induction would be 5.

a. **Percent of Control:** Percent of control is another way to express the relative change in a parameter due to treatment where the obervations range from 0 to a maximum value for both chemical-treated and control samples.

In practice, concentrations are often subjected to multiple testing iterations, resulting in the formation of distinct concentration groups with replicates. The following quantities are introduced corresponding to a single CRS given a compound c_i and assay endpoint a_i:

- $n_{\text{datapoints}_{i,j}}$: the total number of concentration-response pairs ($|S|$)
- $n_{\text{groups}_{i,j}}$: the number of distinct concentrations tested

Figure 4.1: A CRS for the compound *Estropipate* (DTXSID3023005) in the assay endpoint *TOX21_ERa_LUC_VM7_Agonist* (aeid=788). The series has a total of $k = 45$ concentration-response pairs and is composed of $n_{conc} = 15$ concentration groups, each with $n_{rep} = 3$ replicates.

- $n_{\text{replicates}_{i,j}}$: the number of replicates for each concentration group
- $min_{\text{conc}_{i,j}}$: the lowest concentration tested
- $max_{\text{conc}_{i,j}}$: the highest concentration tested

Figure 4.1 showcases a single CRS for some compound tested within an assay endpoint.

To gain a rough visual representation of how these quantities vary across the complete set of analyzed concentration-response series in this thesis, please consult Figure 4.2. This figure illustrates the above metrics aggregated by their means, grouped by assay endpoints and compounds.

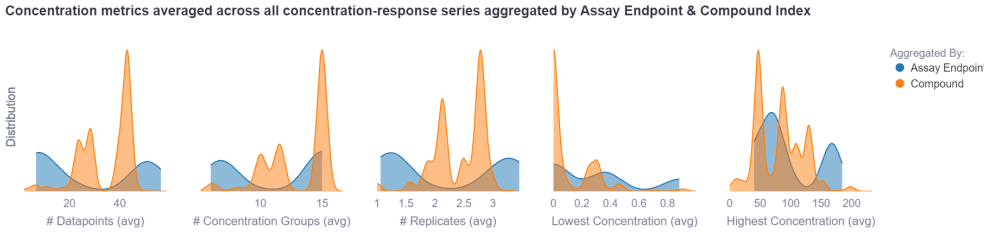


Figure 4.2: Concentration metrics averaged across all concentration-response series aggregated by assay endpoint (blue) and compound (orange). E.g., the first chart shows the distribution on the average number of datapoints across all assay enpoint $a_i \in A$ with $\frac{1}{|C|}\sum_j n_{\text{datapoints}_{i,j}}$ and across all compounds $c_j \in C$ with $\frac{1}{|A|}\sum_i n_{\text{datapoints}_{i,j}}$. Similarly, the process is repeated for the other metrics: $n_{\text{groups}_{i,j}}$, $n_{\text{replicates}_{i,j}}$, $min_{\text{conc}_{i,j}}$, and $max_{\text{conc}_{i,j}}$.

4.2.1 tcplFit2

To improve upon tcpl, `R` package *tcplFit2*[7] was developed, a standalone package focused on curve-fitting and hit-calling. The package also offers a more flexible and robust fitting procedure, allowing for the use of different optimization algorithms and the incorporation of user-defined constraints. *tcplFit2* differs from other `R`-language open-source concentration-response packages like *drc* and *mixtox*, as it is specifically tailored for HTS concentration-response data, offering an extensive set of curve models, summarized in Table 4.1.

Table 4.1: tcplfit2 Model Details

Model	Label	Equations[1]
Constant	cnst	$f(x) = 0$
Linear	poly1	$f(x) = ax$
Quadratic	poly2	$f(x) = a\left(\frac{x}{b} + \left(\frac{x}{b}\right)^2\right)$
Power	pow	$f(x) = ax^p$
Hill	hill	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p}$
Gain-Loss	gnls	$f(x) = \frac{tp}{(1+\left(\frac{ga}{x}\right)^p)(1+\left(\frac{x}{la}\right)^q)}$
Exponential 2	exp2	$f(x) = a\left(\exp\left(\frac{x}{b}\right) - 1\right)$
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$
Exponential 4	exp4	$f(x) = tp\left(1 - 2^{-\frac{x}{ga}}\right)$
Exponential 5	exp5	$f(x) = tp\left(1 - 2^{-\left(\frac{x}{ga}\right)^p}\right)$

[1] Model parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50, la: loss AC50

All the curve fit models assume that the normalized observations from the CRS conform to a Student's t-distribution with 4 degrees of freedom [19] The Student's t-distribution has heavier tails compared to the normal distribution, making it more robust to outlier and eliminates the necessity of removing potential outliers prior to the fitting process. The model fitting algorithm in `tcplFit2` employs nonlinear *maximum likelihood estimation (MLE)* to determine the model parameters for all available models.

[7] https://github.com/USEPA/CompTox-ToxCast-tcplFit2

Consider $t(z, \nu)$ as the Student's t-distribution with ν degrees of freedom, where y_i represents the observed response for the i-th observation, and μ_i is the estimated response for the same observation. The calculation of z_i is as follows:

$$z_i = y_i - \mu_i \exp(\sigma)$$

where σ is the scale term.

Then the log-likelihood is

$$\sum_{i=1}^{n} [\ln(t(z_i, 4)) - \sigma]$$

where n is the number of observations.

The *Akaike Information Criterion (AIC)*[8] is used as as a measure of goodness of fit. The model with the lowest AIC value is chosen as the *winning* model. The winning model is then used to estimate the efficacy and potency of the compound. More precise, the potency estimates, also called *point-of-departure (POD)* estimates, are derived from the fitted curve characteristics, identifying *activity concentrations (AC)* at which the curve crosses certain response levels. For example:

- the AC at which the compound first reaches 50% of its estimated maximum response is denoted by *ac50*
- the AC at which the compound first reaches the estimated efficacy cutoff is denoted by *acc*
- the AC at which the compound first reaches the estimated assay background noise *3bmad*[9] is denoted by *acb*

Figure X illustrates the stated *POD* estimates. Notably, no *POD* estimates are computed when the compound is considered inactive[10], as these estimates are not applicable in such cases.

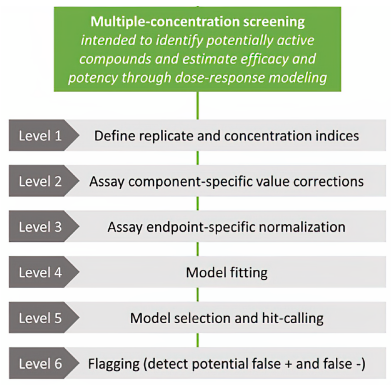
The *continuous hitcall*[11] is calculated based on the product of the probabilities of the following values [18]:

[8] $AIC = -2\log(L(\hat{\theta}, y)) + 2K$

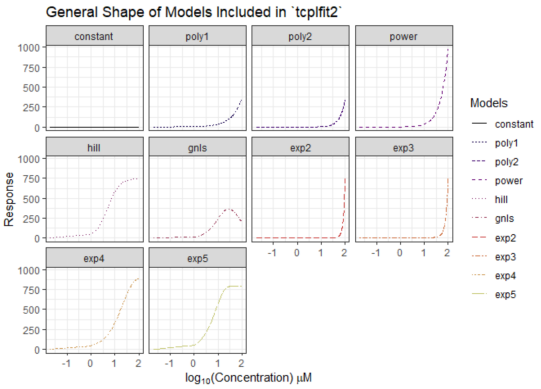
[9] The baseline region is defined as $0 + 3bmad$, where *bmad* represents the median absolute deviation calculated from the response values of the two lowest tested concentrations, where the assumption of the highest level of inactivity is legitimate.

[10] if the *winning* fit model was the constant model

[11] In legacy tcpl only binary hitcall was calculated



(a) Overview of tcpl's data analysis pipeline for multiple concentrations screening data.



(b) Employed curve-fit models in tcpl v3.0 for fitting concentration-response data series through the application of maximum likelihood estimation.

Figure 4.3: tcpl

i. that at least one median response is greater than the efficacy cutoff computed by using the error parameter from the model fit and Student t-distribution to calculate the odds of at least one response exceeding the efficacy cutoff

ii. that the top of the winning fitted curve is above the cutoff:
the likelihood ratio of the one-sided probability of the efficacy cutoff being exceeded

iii. that the winning AIC value is less than that of the constant model:

$$\frac{e^{-\frac{1}{2}AIC_{\text{winning}}}}{e^{-\frac{1}{2}AIC_{\text{winning}}} + e^{-\frac{1}{2}AIC_{\text{cnst}}}}$$

Finally, after processing, each CRS is categorized into an appropriate fit category based on the level of certainty in the estimated bioactivity. Additionally, cautionary flags are assigned to account for problematic data series or uncertain fits and hits.

4.3 Data Overview

Presence Matrix

To enhance data comprehension, we introduce a presence matrix denoted as $P \in 0, 1^{m \times n}$. In this matrix, rows (indexed by i) represent assay endpoints

a_i, and columns (indexed by j) indicate whether testing was conducted[12] (1) or not conducted (0) for compound c_j in those endpoints. Due to selective compound testing across different assay endpoints, matrix P is sparse.

For a visual representation of the presence matrix P covering all assay endpoints and compounds in *invitroDBv4.1*, refer to Figure 4.4.

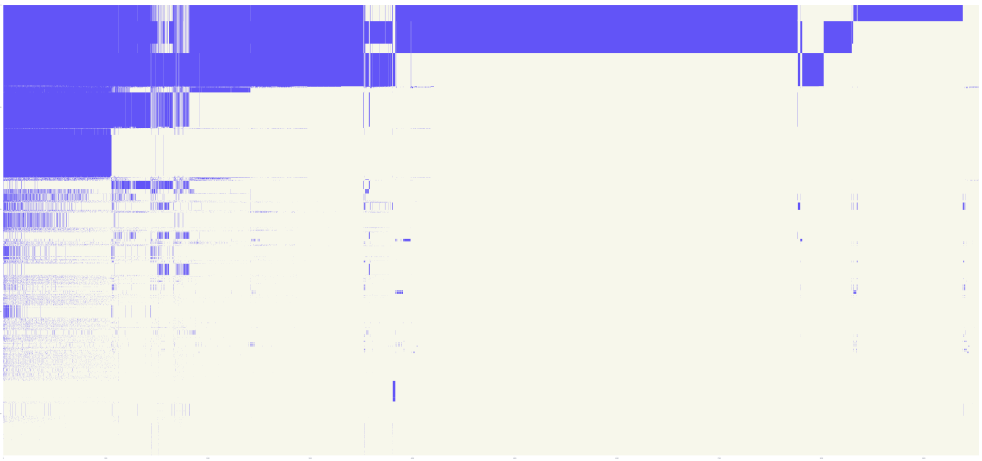


Figure 4.4: The *presence matrix* P covering all assay endpoints and compounds available in *invitroDBv3.5* with $m = 2205$ assay endpoints and $n = 9541$ compounds. The presence matrix is organized by sorting it based on the number of compounds present in each assay endpoint and the compounds are arranged in descending order of their presence frequency. The total count, where $P_{ij} = 1$, indicates the availability of 3 342 377 concentration-response series for downstream analysis.

Subsetting data

For this thesis only assay endpoints that have been tested with a minimum of 1000 compounds were exclusively taken into account. This criterion ensures the availability of sufficient data to train a machine learning model with a minimum level of robustness. Please consult Figure 4.5 to view a graphical depiction of the presence matrix P, which includes only the specific considerd subset of assay endpoints. From now on, we will call this specific subset the *data* that we will be focusing on for this thesis.

We analyse in total $\sum_{i,j} P_{ij} = 1\,372\,225$ concentration-response series, comprising a sum of $\sum_{i,j} |S_{ij}| = 48\,861\,036$ concentration-response pairs across all compounds and assay endpoints.

[12] A compound is regarded as having been tested within an assay endpoint when there exists a corresponding concentration-response series accessible in the database.

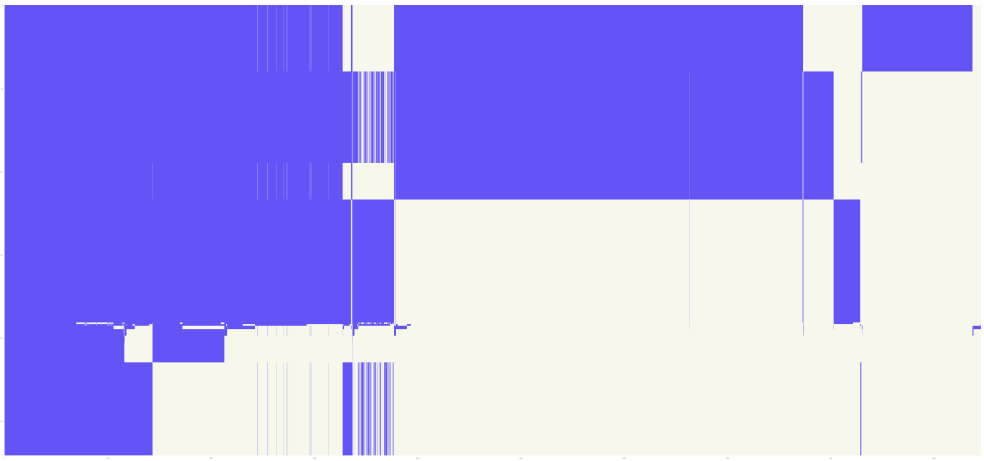


Figure 4.5: The *presence matrix* P covering only the subset of all of assay endpoints available in *invitroDBv3.5*, considered for this thesis, encompassing $m = 271$ assay endpoints and $n = 9456$ compounds. The total count, where $P_{ij} = 1$, indicates the availability of 1 372 225 concentration-response series for downstream analysis.

4.4 pytcpl

This thesis introduces *pytcpl*[13], a streamlined `Python` package inspired by the `R` packages *tcpl* and tcplfit2, which were extensively discussed in Chapter 2. The package optimizes data storage and generates compressed *Parquet* files of the relevant raw data and metadata from *invitroDBv4.1*. This efficient strategy reduces storage needs, resulting in less than 10 GB within the repository, compared to the original 80 GB database. This obviates the need for a cumbersome, large-scale database installation, rendering downstream analysis more accessible and efficient. Our package is crafted to accomodate cusomizable processing steps and facilitate interactive data visualization with a novel *Curve Surfer*[14]. Furthermore, it enables researchers who prefer `Python` to easily participate in data analysis and exploration, overcoming any limitations associated with using `R` code.

4.4.1 Pipeline

The *pytcpl* pipeline is composed of four main steps, as illustrated in Figure **??**, summarising the steps also done by tcpl and tcplfit2.

1. Data collection
2. Cutoff determination and filtering (Meet conditions for curve fitting)

[13] https://github.com/rbBosshard/pytcpl
[14] https://pytcpl.streamlit.app/

3. Curve fitting
4. Hit calling

Model	Label	Equations[1]	Role in pytcpl
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$	Omit
Gain-Loss 2	gnls2	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p} \exp\left(-qx\right)$	New

[1] Model parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50

Table 4.2: tcplfit2 Model Details

4.4.2 Curve Surfer

Data visualization, overview of what is possible with the tool. Filter by assay endpoint, compound, etc.

4.5 Machine Learning Pipeline

4.5.1 Preprocessing

Subselecting the columns from the output tables generated by pytcpl: DTXSID identifier and continuous hitcall value. The feature inputs to the machine learning model is a molecular structure represented as fingerprint generated from a SMILES string uniquely determined by the compounds DTXSID identifier. The SMILES string is a linear representation of a compound's molecular structure. The SMILES string is converted to a molecular graph, which is then converted to a feature vector. The feature vector is then used to train a machine learning model. The machine learning model is then used to predict the hitcall value for a given compound. The machine learning pipeline is illustrated in Figure?.

4.5.2 Binary Classification

The goal is to predict whether a compound is active or inactive for a given assay endpoint. We can formulate this as a binary classification problem, where the input is the compound's molecular structure fingerprint and the output is the hitcall value binarized by some decision threshold. The hitcall value is rendered to a binary variable, where 1 indicates that the compound is active and 0 indicates that the compound is inactive.

4.5.3 Regression

4.5.4 Massbank Validation

Chapter 5

Results and Discussion

5.1 Results

5.2 Evaluation

5.3 Discussion

5.4 Performance Analysis

5.5 Binary Classification

When evaluating the effectiveness of a binary prediction technique using a validation dataset with known activities, four central values are considered:

1. True Positives (TP): The number of correctly predicted active cases.
2. True Negatives (TN): The number of correctly predicted inactive cases.
3. False Positives (FP): The number of incorrectly predicted active cases.
4. False Negatives (FN): The number of incorrectly predicted inactive cases.

Chapter 6

Conclusion

We have evidence of a multitude of chemicals being present in the environment and in our bodies and that mixture exposure indeed matters. This knowledge needs to be deepened, and the quantitative contribution of chemicals to compromised health should be better described and translated into regulatory action. As indicated in a scientific opinion paper of the German Federal Environmental Agency (Conrad et al. 2021), the CSS goals may be considered as a moving target. For increasing scientific evidence and improved method for detection and assessment of chemicals, development of new technologies require innovative regulatory, technological and societal reactions. We should be flexible and prepared to take up the scientific challenges and collaborate productively with regulatory institutions to address the identified challenges and modernise chemical risk assessment. This is also in line with the concern of many scientists that chemical pollution and the wide range of adverse effects on human and ecosystem health demand additional efforts on a global scale (Brack et al. 2022; Wang et al. 2021). We see the CSS as a European strategy that, in concert with other initiatives, may open new opportunities to minimise hazardous chemical pollution and thus risks to human health and ecosystems.

6.1 Further

mistures are not combination of mixtures toxic effects are not tested and produce an exponential times more assessment work

Bibliography

[1] U. N. E. Programme, *Global chemicals outlook ii - from legacies to innovative solutions: Implementing the 2030 agenda for sustainable development - synthesis report*, 2019. [Online]. Available: https://wedocs.unep.org/20.500.11822/27651.

[2] C. A. Service, *Chemical abstracts service (cas) is a division of the american chemical society*, Source of chemical information located in Columbus, Ohio, United States, https://www.cas.org/support/documentation/cas-databases, 2023.

[3] R. Schwarzenbach *et al.*, "The challenge of micropollutants in aquatic systems," *Science (New York, N.Y.)*, vol. 313, pp. 1072–7, Sep. 2006. DOI: 10.1126/science.1127291.

[4] E. Commission, D.-G. for Research, and Innovation, *European Green Deal - Research & innovation call*. Publications Office of the European Union, 2021. DOI: 10.2777/33415.

[5] E. Commission, "Eu chemicals strategy for sustainability towards a toxic-free environment," 2020, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability Towards a Toxic-Free Environment. [Online]. Available: https://environment.ec.europa.eu/strategy/chemicals-strategy_en.

[6] J. Hollender, E. L. Schymanski, H. P. Singer, and P. L. Ferguson, "Non-target screening with high resolution mass spectrometry in the environment: Ready to go?" *Environmental Science & Technology*, vol. 51, no. 20, pp. 11 505–11 512, 2017, PMID: 28877430. DOI: 10.1021/acs.est.7b02184. eprint: https://doi.org/10.1021/acs.est.7b02184. [Online]. Available: https://doi.org/10.1021/acs.est.7b02184.

[7] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: a webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research*, vol. 46, no. W1, W257–W263, Apr. 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky318. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W257/25110434/gky318.pdf. [Online]. Available: https://doi.org/10.1093/nar/gky318.

[8] N. H. G. R. I. Maggie Bartlett. "Chemical genomics robot." (2009), [Online]. Available: https://en.wikipedia.org/wiki/High-throughput_screening#/media/File:Chemical_Genomics_Robot.jpg.

[9] J. Rudd. "High throughput screening - accelerating drug discovery efforts." (2017), [Online]. Available: https://www.ddw-online.com/hts-a-strategy-for-drug-discovery-900-200008/.

[10] K. Arturi and J. Hollender, "Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data," *Environmental Science & Technology*, vol. 0, no. 0, null, 0, PMID: 37279189. DOI: 10.1021/acs.est.3c00304. eprint: https://doi.org/10.1021/acs.est.3c00304. [Online]. Available: https://doi.org/10.1021/acs.est.3c00304.

[11] K. Dührkop *et al.*, "Sirius 4: A rapid tool for turning tandem mass spectra into metabolite structure information," *Nature methods*, vol. 16, no. 4, pp. 299–302, Apr. 2019, ISSN: 1548-7091. DOI: 10.1038/s41592-019-0344-8. [Online]. Available: https://research.aalto.fi/files/32997691/SCI_Duhrkop_Fleischauer_Sirius_4_Turning_tandem.pdf.

[12] *Massbank: High quality mass spectral database*, https://massbank.eu/MassBank/, Accessed: 2023.

[13] T. Janela, K. Takeuchi, and J. Bajorath, "Introducing a chemically intuitive core-substituent fingerprint designed to explore structural requirements for effective similarity searching and machine learning," *Molecules*, vol. 27, no. 7, 2022, ISSN: 1420-3049. DOI: 10.3390/molecules27072331. [Online]. Available: https://www.mdpi.com/1420-3049/27/7/2331.

[14] S. M. Bell *et al.*, "In vitro to in vivo extrapolation for high throughput prioritization and decision making," *Toxicology in Vitro*, vol. 47, pp. 213–227, 2018, ISSN: 0887-2333. DOI: https://doi.org/10.1016/j.tiv.2017.11.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0887233317303661.

[15] P. Nymark *et al.*, "Systematic organization of covid-19 data supported by the adverse outcome pathway framework," *Frontiers in Public Health*, vol. 9, May 2021. DOI: 10.3389/fpubh.2021.638605.

[16] R. Judson *et al.*, "Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space," *Toxicological Sciences*, vol. 152, no. 2, pp. 323–339, May 2016, ISSN: 1096-6080. DOI: 10.1093/toxsci/kfw092. eprint: https://academic.oup.com/toxsci/article-pdf/152/2/323/26290632/kfw092.pdf. [Online]. Available: https://doi.org/10.1093/toxsci/kfw092.

[17] B. Escher, P. Neale, and F. Leusch, *Bioanalytical Tools in Water Quality Assessment*. IWA Publishing, Jun. 2021, ISBN: 9781789061987. DOI: 10.2166/9781789061987. eprint: https://iwaponline.com/book-pdf/899726/wio9781789061987.pdf. [Online]. Available: https://doi.org/10.2166/9781789061987.

[18] T. Sheffield, J. Brown, S. Davidson, K. P. Friedman, and R. Judson, "tcplfit2: an R-language general purpose concentration–response modeling package," *Bioinformatics*, vol. 38, no. 4, pp. 1157–1158, Nov. 2021, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab779. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/4/1157/50422999/btab779.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab779.

[19] C. for Computational Toxicology and U. E. Exposure, *Tcpl v3.0 data processing*, R package vignette for the tcpl package v3.0, CRAN, 2023. [Online]. Available: https://cran.r-project.org/web/packages/tcpl/vignettes/Data_processing.html.